

Non asymptotic distributional bounds for the Dickman Approximation of the running time of the Quickselect algorithm

Larry Goldstein

March 3, 2017

Abstract

Given a non-negative random variable W and $\theta > 0$, let the generalized Dickman transformation map the distribution of W to that of

$$W^* =_d U^{1/\theta}(W + 1),$$

where $U \sim \mathcal{U}[0, 1]$, a uniformly distributed variable on the unit interval, independent of W , where $=_d$ denotes equality in distribution. It is well known that W^* and W are equal in distribution if and only if W has the generalized Dickman distribution \mathcal{D}_θ . By use of Stein's method, we demonstrate that the Wasserstein distance d_1 between W , a non-negative random variable with finite mean, and D_θ having distribution \mathcal{D}_θ obeys the inequality

$$d_1(W, D_\theta) \leq (1 + \theta)d_1(W, W^*).$$

The specialization of this bound to the case $\theta = 1$ and coupling constructions yield

$$d_1(W_n, D) \leq \frac{8 \log(en/2) + 2}{n} \quad \text{for all } n \geq 1, \text{ where } W_n = \frac{1}{n}C_n - 1,$$

with C_n the number of comparisons made by the Quickselect algorithm to find the smallest element of a list of n distinct numbers. Results for the running time for Quickselect to locate the m^{th} smallest element of the list obey a similar bound, and in particular recover the results of [12] that show distributional convergence of W_n to the standard Dickman distribution in the asymptotic regime $m = o(n)$.

1 Introduction

For a given non-negative random variable W , let the generalized Dickman transformation map the distribution of W to that of W^* satisfying

$$W^* =_d U^{1/\theta}(W + 1) \tag{1}$$

⁰This work was partially supported by NSA grant H98230-15-1-0250.

⁰MSC 2010 subject classifications: Primary 60F05, 68Q25

⁰Key words and phrases: sorting, complexity, integral equation, distributional approximation

where $U \sim \mathcal{U}[0, 1]$, a uniformly distributed variable on the unit interval, independent of W and where $=_d$ denotes equality in distribution. It is well known [7], [16] that the generalized Dickman distribution \mathcal{D}_θ is the unique fixed point of the transformation (1), that is,

$$W \sim \mathcal{D}_\theta \quad \text{if and only if} \quad W =_d W^*. \quad (2)$$

When (1) holds we will say that W^* has the \mathcal{D}_θ -bias distribution of W . In what follows, D_θ will denote a random variable with distribution \mathcal{D}_θ . The case $\theta = 1$ corresponds to the (standard) Dickman distribution, for which we will drop the subscript θ , denoting, for instance, D_1 by D .

The Dickman distribution first made its appearance in number theory [8] when counting the number of integers below a fixed threshold whose prime factors satisfy some given upper bound; see also the more recent work [14] in this area. Members from the generalized Dickman family have subsequently been noted to arise in a variety of other contexts, in particular, for component counts of logarithmic combinatorial structures such as permutations and partitions in [3], and more generally for the quasi-logarithmic class by considered in [4], for the sum of edge lengths of vertices connected to the origin in minimal directed spanning trees in [16], and for certain weighted sums of independent random variables [15], such as that in (9). Simulation of the Dickman distribution has been considered in [7].

Here we study the Dickman approximation for the limiting distribution of the Quickselect sorting algorithm of Hoare [11] for locating the m^{th} smallest element of a list of n distinct numbers. One may visualize how Quickselect works in terms of a tree structure. First, a number from the given list is chosen uniformly. This value is used as a ‘key’ to sort the remaining values into those that are strictly smaller, making up the left subtree, and those that are strictly larger, making up the right. If the left subtree is of size $m - 1$ then the key is the desired m^{th} smallest element, and the procedure terminates. Otherwise, the process continues recursively on the left sub-tree if it is of size m or larger, and else on the right sub-tree.

Letting

$$W_{n,m} = \frac{1}{n} C_{n,m} - 1, \quad (3)$$

where $C_{n,m}$ is the number of comparisons made by Quickselect, the work of [12] showed that $W_{n,m}$ converges in distribution to the Dickman D when $m = o(n)$. We note that in the case $m = 1$ Quickselect simplifies in that at each step of the recursion the procedure either stops or continues on the left subtree. As this case is simpler to handle than when $m \geq 2$, we deal with these separately. The following two theorems quantify and recover the results of [12] by providing non-asymptotic bounds in the Wasserstein distance d_1 between $W_{n,m}$ and D that converge to zero in the $m = o(n)$ asymptotic regime. Noting that Theorem 1.1 handles the case $m = 1$, the restriction that $n \geq m$ made in both the following two results is imposed only to enforce non-triviality, that is, only to assure that the m^{th} smallest element of a list of n distinct numbers exists.

Theorem 1.1 *Let $C_{n,1}$ be the number of comparisons made by Quickselect to find the smallest of a list of n distinct numbers, and let $W_{n,1}$ be given by (3). Then for all $n \geq 1$*

$$d_1(W_{n,1}, D) \leq \frac{8 \log(en/2) + 2}{n}.$$

Theorem 1.2 *Let $m \geq 2$ and $C_{n,m}$ the number of comparisons made by Quickselect to find the m^{th} smallest element of a list of n distinct numbers, and let $W_{n,m}$ be given by (3). Then for all $n \geq m$*

$$d_1(W_{n,m}, D) \leq \frac{(8 + 46m) \log(en/m) + 8m}{n}.$$

Theorems 1.1 and 1.2 are derived by applying Theorem 1.3 that quantifies the if direction of the fixed point property (2) in the Wasserstein, or d_1 metric between two random variables X Y , given by

$$d_1(X, Y) = \sup_{h \in \text{Lip}_1} |Eh(X) - Eh(Y)|, \quad (4)$$

where for any $\alpha \geq 0$

$$\text{Lip}_\alpha = \{h : |h(y) - h(x)| \leq \alpha|y - x|\}. \quad (5)$$

On the left hand side of (4) we have chosen to write $d_1(X, Y)$ rather than the technically correct expression $d_1(\mathcal{L}(X), \mathcal{L}(Y))$ only for notational convenience.

Theorem 1.3 *Let W be a non-negative random variable with finite mean, $\theta > 0$, and let the law of W^* be given by (1). Then*

$$d_1(W, D_\theta) \leq (1 + \theta)d_1(W^*, W). \quad (6)$$

Theorem 1.3 is obtained by developing Stein's method for the case of the generalized Dickman. Stein's method originated in the seminal works [17] and [18] for the normal, see also the extensive and more modern treatment [5]. The extension of Stein's method to this realm is somewhat non-standard, as the Stein equations here are of a differential-delay and integral type, see (44) and (45) respectively, for which bounds on the solution are not readily available via standard tools. Nevertheless, the bound (6) parallels the one of the same form in the normal case given in Theorem 1.1 of [9], also shown using Stein's method, having constant of 2 on the right hand side, with W^* there indicating the zero bias transform having unique fixed point at the normal.

As the Wasserstein distance also satisfies (see [13], for instance)

$$d_1(X, Y) = \inf E|X - Y| \quad (7)$$

where the infimum is over all couplings (X, Y) having the given marginals, Theorem 1.3 implies that

$$d_1(W, D_\theta) \leq (1 + \theta)E|W^* - W| \quad (8)$$

for any non-negative random variables W and W^* defined on a common space, with W^* having the \mathcal{D}_θ -bias distribution of W .

In related work, [6] obtains distributional bounds for the running time of a variation of Quickselect to a non-Dickman approximand, in particular compare its characterization in (1.4) there to (1) here. Further, [1] considers the Dickman approximation of sums of the form

$$W_n = \frac{1}{n} \sum_{k=1}^n kY_k \quad (9)$$

where $Y_k, k = 1, 2, \dots$ are independent random Bernoulli random variables with success probabilities given by $P(Y_k = k) = 1/k$. Their approach is based on the rewriting of (39) in the case $\theta = 1$, with f' replaced by f , to obtain that the standard Dickman is characterized by the size bias type relation

$$E[Df(D)] = E[f(D + U)] \quad (10)$$

where $U \sim \mathcal{U}[0, 1]$, and is independent of D ; for the use of size biasing in general, and in Stein's method in particular, see [2], [10] and [5]. The authors of [1] then use (10) and coupling methods to obtain a bound on the difference between the characteristic functions of D and W_n in (9), and through that obtain a bound of the form $C\sqrt{\log n}/n$ on a weakened Wasserstein metric obtained by restricting the Lip_1 class in (4) to three times differentiable test functions. In particular, the work [1] does not consider any form of the Stein equation, such as those in (44) or (45), and consequently do not obtain bounds on its solution for any Dickman case, as is achieved here in Theorem 3.1. Indeed they note in general that this last step can be an 'extremely difficult problem'.

In Section 2 we detail the workings of the Quickselect algorithm and prove Theorems 1.1 and 1.2 by applying Theorem 1.3. In Section 3 we explore the properties of a certain averaging operator that arises in our analysis, set up the Stein equation for the Dickman through its use, and prove Theorem 1.3 using the tools developed. One technical challenge that appears in the development is obtaining bounds on the smoothness of a particular solution to a certain integral equation. This matter is resolved in Theorem 3.1.

2 Quickselect

In this section we apply Theorem 1.3 to obtain the error bounds in Theorems 1.1 and 1.2 on the distribution of the running time C_{nm} of the Quickselect algorithm for finding the m^{th} smallest element of a list of n distinct numbers. When the value of m is clear from context, we will write C_n for $C_{n,m}$.

2.1 Quickselect: the case $m = 1$

In this section we prove Theorem 1.1 for the distribution of the number of comparisons C_n that Quickselect requires to locate the smallest element of a list of n distinct numbers. Clearly, an empty list requires no comparisons, hence $C_0 = 0$. For $n \geq 1$, the procedure requires the $n - 1$ comparisons of the chosen key to every other element at the first stage, followed by the cost of processing the left subtree, which may be empty. Since the key is chosen uniformly, we obtain the stochastic recursion

$$C_n = n - 1 + C_{V_1} \quad \text{for } n \geq 1, \text{ with boundary condition } C_0 = 0, \quad (11)$$

where V_1 , the size of the left subtree, is a discrete uniform variable on $\{0, \dots, n - 1\}$. From (11) we see that $C_1 = 0$ and $C_2 = 1$ a.s., and that non-trivial distributions arise for $n \geq 3$.

Before proceeding to the proof of the theorem we describe how for all $n \geq 1$ we may write C_n as a function $C(n; \mathbf{U}_1)$ of n and

$$\mathbf{U}_k = (U_k, U_{k+1}, \dots) \quad \text{for } k \geq 1,$$

where U_1, U_2, \dots is a sequence of i.i.d. uniform variables on $[0, 1]$. Consider the initial list of size $V_0 = n$ as making up the left subtree for stage 0. At stage $k \geq 1$, given a non-null left subtree from the previous stage of size V_{k-1} , the new left subtree results by choosing a key uniformly from the current left subtree, resulting in one of size of size

$$V_k = \lfloor V_{k-1} U_k \rfloor \quad \text{for } k \geq 1, \quad (12)$$

satisfying $V_k \sim \mathcal{U}\{0, \dots, V_{k-1} - 1\}$. Rewriting (11) in this notation we have

$$C(n; \mathbf{U}_1) = n - 1 + C(\lfloor n U_1 \rfloor; \mathbf{U}_2) \quad \forall n \geq 1, \text{ with } C(0; \mathbf{U}_k) = 0 \quad \forall k \geq 1. \quad (13)$$

As the size of each non-null left subtree decrements by at least one at each iteration, the value of C_n will only depend on an initial subsequence of \mathbf{U}_1 of length at most n .

We pause to prove the following lemma that is needed in this section, and the one following.

Lemma 2.1 *If for c a non-negative number and m a positive integer*

$$e_n \leq c + \frac{1}{n} \sum_{u=m}^{n-1} e_u \quad \text{for all } n \geq m, \quad (14)$$

then

$$e_n \leq c \log(en/m) \quad \text{for } n \geq m. \quad (15)$$

Proof: As (14) holds for $n = m$ we see that $e_m \leq c$, verifying that the inequality in (15) holds at m . Assuming inequality (14) holds for $m \leq u \leq n - 1$ for some $n \geq m + 1$ we have

$$\begin{aligned} e_n &\leq c + \frac{1}{n} \sum_{u=m}^{n-1} e_u \leq c + \frac{c}{n} \sum_{u=m}^{n-1} \log(eu/m) \leq c + \frac{c}{n} \int_m^n \log(eu/m) du \\ &= c \left(1 + \frac{1}{n} \left[u \log(eu/m) - u \right]_m^n \right) = c \left(1 + \frac{1}{n} [n \log(en/m) - n] \right) = c \log(en/m), \end{aligned}$$

completing the inductive step, and the proof of the lemma. \square

We now prove Theorem 1.1, deferring the proofs of Lemmas 2.2 and 2.4 to the end of this section.

Proof of Theorem 1.1: Take $n \geq 1$. With V_k as in (12), by (13) the variable W_n as given by (3) satisfies

$$W_n = \frac{1}{n} C(n; \mathbf{U}_1) - 1 = \frac{1}{n} (n - 1 + C(V_1; \mathbf{U}_2)) - 1 = \frac{1}{n} (C(V_1; \mathbf{U}_2) - 1).$$

We now construct a variable with the W_n^* distribution by first constructing a W'_n having the W_n distribution. As \mathbf{U}_1 and \mathbf{U}_2 are equidistributed,

$$W'_n := \frac{1}{n} C(n, \mathbf{U}_2) - 1 =_d \frac{1}{n} C(n, \mathbf{U}_1) - 1 = W_n,$$

and hence

$$W_n^* = U_1(W_n' + 1) = \frac{1}{n}U_1C(n; \mathbf{U}_2) \quad (16)$$

has the \mathcal{D} -bias distribution by (1). Taking the difference,

$$W_n^* - W_n = \frac{1}{n}(U_1C(n; \mathbf{U}_2) - C(V_1; \mathbf{U}_2) + 1),$$

for which

$$nE|W_n^* - W_n| \leq e_n + 1, \quad \text{where we set } e_k = E|U_1C(k; \mathbf{U}_2) - C(\lfloor kU_1 \rfloor; \mathbf{U}_2)|, \quad k \geq 0.$$

Consequence (8) of Theorem 1.3 with $\theta = 1$ yields

$$d_1(W_n, D) \leq 2E|W_n^* - W_n| = \frac{2}{n}(e_n + 1). \quad (17)$$

We claim that

$$\begin{aligned} e_n &= E|U_1C(n; \mathbf{U}_2) - C(\lfloor nU_1 \rfloor; \mathbf{U}_2)| \\ &\leq E|U_1(n-1) - \lfloor nU_1 \rfloor + 1| + E|U_1C(\lfloor nU_2 \rfloor; \mathbf{U}_3) - C(\lfloor \lfloor nU_1 \rfloor U_2 \rfloor; \mathbf{U}_3)|. \end{aligned}$$

When $\lfloor nU_1 \rfloor \geq 1$ this inequality follows from using the basic recursion (13) to both terms followed by applying the triangle inequality, and is easily verified to hold directly in the case $\lfloor nU_1 \rfloor = 0$ by applying (13) only on the first term, noting the second one in this case is zero. Now using that $|u(n-1) - \lfloor nu \rfloor + 1| \leq 2$ for all $u \in [0, 1]$, we obtain

$$\begin{aligned} e_n &\leq 2 + E|U_1C(\lfloor nU_2 \rfloor; \mathbf{U}_3) - C(\lfloor \lfloor nU_1 \rfloor U_2 \rfloor; \mathbf{U}_3)| \\ &\leq 2 + E|U_1C(\lfloor nU_2 \rfloor; \mathbf{U}_3) - C(\lfloor \lfloor nU_2 \rfloor U_1 \rfloor; \mathbf{U}_3) + 1| + E|C(\lfloor \lfloor nU_2 \rfloor U_1 \rfloor; \mathbf{U}_3) - C(\lfloor \lfloor nU_1 \rfloor U_2 \rfloor; \mathbf{U}_3)| \\ &= 2 + Ee_{\lfloor nU_2 \rfloor} + E|C(\lfloor \lfloor nU_2 \rfloor U_1 \rfloor; \mathbf{U}_3) - C(\lfloor \lfloor nU_1 \rfloor U_2 \rfloor; \mathbf{U}_3)| \\ &\leq 4 + Ee_{\lfloor nU_2 \rfloor}, \quad (18) \end{aligned}$$

where for the final term, the inequality

$$E|C(\lfloor \lfloor nU_2 \rfloor U_1 \rfloor; \mathbf{U}_3) - C(\lfloor \lfloor nU_1 \rfloor U_2 \rfloor; \mathbf{U}_3)| \leq 2 \quad \text{for all } n \geq 0$$

follows by applying Lemma 2.2, below, that shows that $|\lfloor U_1 \lfloor nU_2 \rfloor \rfloor - \lfloor U_2 \lfloor nU_1 \rfloor \rfloor| \leq 1$ a.s, and Lemma 2.4, also below, that shows that $E|C(p, \mathbf{U}_3) - C(p-1, \mathbf{U}_3)| \leq 2$ for all $p \geq 1$.

Expanding the expectation in $Ee_{\lfloor nU_2 \rfloor}$ in (18), using that $\lfloor nU_2 \rfloor$ is uniformly distributed over $\{0, \dots, n-1\}$ and that $e_0 = e_1 = 0$ by virtue of $C_0 = C_1 = 0$, we obtain

$$e_n \leq 4 + \frac{1}{n} \sum_{u=0}^{n-1} e_u \leq 4 + \frac{1}{n} \sum_{u=2}^{n-1} e_u \quad \text{for } n \geq 2.$$

As $e_1 = 0$ inequality (17) shows that the claim of the theorem holds for $n = 1$. Applying Lemma 2.1 with $m = 2$ shows that $e_n \leq 4 \log(en/2)$ for $n \geq 2$, and substituting this bound into (17) now completes the proof. \square

We now prove Lemmas 2.2 and 2.4.

Lemma 2.2 For all $\{u_1, u_2\} \subset [0, 1)$ and $n \geq 0$,

$$||u_1 \lfloor nu_2 \rfloor - u_2 \lfloor nu_1 \rfloor|| \leq 1.$$

Proof: Consider the case $n \geq 1$, else the claim is trivial. Let $s = \lfloor nu_1 \rfloor$ and $t = \lfloor nu_2 \rfloor$, so that $\{s, t\} \subset \{0, 1, \dots, n-1\}$ and

$$s \leq nu_1 < (s+1) \quad \text{and} \quad t \leq nu_2 < (t+1).$$

Then

$$\frac{st}{n} \leq u_2 \lfloor nu_1 \rfloor < \frac{s(t+1)}{n} \quad \text{and} \quad \frac{st}{n} \leq u_1 \lfloor nu_2 \rfloor < \frac{(s+1)t}{n}.$$

Taking the difference,

$$|u_1 \lfloor nu_2 \rfloor - u_2 \lfloor nu_1 \rfloor| < \frac{1}{n} \max\{s, t\} < 1.$$

As the difference between $u_1 \lfloor nu_2 \rfloor$ and $u_2 \lfloor nu_1 \rfloor$ is less than 1, their integer parts can differ by at most 1. \square

One may easily verify that

$$\frac{k-1}{p-1} < \frac{k}{p} < \frac{k}{p-1} \quad \text{for } p \geq 2 \text{ and } 1 \leq k \leq p-1, \quad (19)$$

and for $u \in [0, 1]$,

$$(\lfloor (p-1)u \rfloor, \lfloor pu \rfloor) = \begin{cases} (k-1, k-1) & u \in \left[\frac{k-1}{p-1}, \frac{k}{p}\right) \\ (k-1, k) & u \in \left[\frac{k}{p}, \frac{k}{p-1}\right). \end{cases} \quad (20)$$

The following fact can be shown directly using induction.

Lemma 2.3 If $c \geq 0$, $f_1 = 0$ and

$$f_p \leq c + \frac{1}{p(p-1)} \sum_{k=1}^{p-1} k f_k \quad \text{for all } p \geq 2$$

then $f_p \leq 2c$ for all $p \geq 1$.

Lemma 2.4 For all $p \geq 1$

$$f_p := E|C(p, \mathbf{U}_1) - C(p-1, \mathbf{U}_1)| \leq 2.$$

Proof: As $f_1 = 0$ we need only consider $p \geq 2$. In view of (19) we may write

$$\begin{aligned} f_p &= E|C(p, \mathbf{U}_1) - C(p-1, \mathbf{U}_1)| \\ &= \sum_{k=1}^{p-1} E \left[|C(p, \mathbf{U}_1) - C(p-1, \mathbf{U}_1)| \mid U_1 \in \left[\frac{k-1}{p-1}, \frac{k}{p}\right) \right] P \left(U_1 \in \left[\frac{k-1}{p-1}, \frac{k}{p}\right) \right) \\ &\quad + \sum_{k=1}^{p-1} E \left[|C(p, \mathbf{U}_1) - C(p-1, \mathbf{U}_1)| \mid U_1 \in \left[\frac{k}{p}, \frac{k}{p-1}\right) \right] P \left(U_1 \in \left[\frac{k}{p}, \frac{k}{p-1}\right) \right). \end{aligned}$$

We claim that the conditional expectation in the first sum is 1. Indeed, given that $U_1 \in [(k-1)/(p-1), k/p)$, the first case of (20) yields $(\lfloor (p-1)U_1 \rfloor, \lfloor pU_1 \rfloor) = (k-1, k-1)$, and now (13) implies that on this event

$$C(p, \mathbf{U}_1) - C(p-1, \mathbf{U}_1) = p-1 + C(k, \mathbf{U}_2) - (p-2 + C(k, \mathbf{U}_2)) = 1.$$

For the second sum, the second case of (20) yields $(\lfloor (p-1)U_1 \rfloor, \lfloor pU_1 \rfloor) = (k-1, k)$, and

$$\begin{aligned} C(p, \mathbf{U}_1) - C(p-1, \mathbf{U}_1) &= p-1 + C(k, \mathbf{U}_2) - (p-2 + C(k-1, \mathbf{U}_2)) \\ &= 1 + C(k, \mathbf{U}_2) - C(k-1, \mathbf{U}_2). \end{aligned}$$

Hence,

$$\begin{aligned} f_p &= \sum_{k=1}^{p-1} P\left(U_1 \in \left[\frac{k-1}{p-1}, \frac{k}{p}\right)\right) \\ &\quad + \sum_{k=1}^{p-1} E\left[\left|1 + C(k, \mathbf{U}_2) - C(k-1, \mathbf{U}_2)\right| \mid U_1 \in \left[\frac{k}{p}, \frac{k}{p-1}\right]\right] P\left(U_1 \in \left[\frac{k}{p}, \frac{k}{p-1}\right)\right) \\ &\leq \sum_{k=1}^{p-1} P\left(U_1 \in \left[\frac{k-1}{p-1}, \frac{k}{p}\right)\right) + \sum_{k=1}^{p-1} (1 + f_k) P\left(U_1 \in \left[\frac{k}{p}, \frac{k}{p-1}\right)\right) \\ &= 1 + \sum_{k=1}^{p-1} f_k P\left(U_1 \in \left[\frac{k}{p}, \frac{k}{p-1}\right)\right) = 1 + \frac{1}{p(p-1)} \sum_{k=1}^{p-1} k f_k. \end{aligned}$$

Invoking Lemma 2.3 with $c = 1$ now completes the proof. \square

Though the following lemma will only be used in our study of the case $m \geq 2$ in Subsection 2.2, as it pertains to the case $m = 1$ we place it in the current section.

Lemma 2.5 *The expected running time of the Quickselect algorithm for $m = 1$ obeys the bound $E[C(n, \mathbf{U}_1)] \leq 2n$ for all $n \geq 0$.*

Proof: Set $a_n = E[C(n, \mathbf{U}_1)]$ for all $n \geq 0$ and note that clearly $a_0 = 0$. Now, for $n \geq 1$, taking expectation in (13) and using that $V_1 \sim \mathcal{U}\{0, \dots, n-1\}$ yields

$$a_n \leq n-1 + \frac{1}{n} \sum_{k=0}^{n-1} a_k \quad \text{for } n \geq 1.$$

The claim now follows immediately by induction. \square

2.2 Case of $m \geq 2$

In this section we prove Theorem 1.2, handling the case where for $m \geq 2$ we are searching for the m^{th} smallest element of a list of n distinct numbers. When $n \leq m-1$ the m^{th} smallest element of the list does not exist, so in this trivial case no comparison are required. Otherwise, the Quickselect algorithm, say Q_m , begins as for $m = 1$ at the first stage by selecting a uniformly chosen key, giving rise to a left subtree of size V_1 , uniformly distributed in $\{0, \dots, n-1\}$ and a right subtree of size $n-1-V_1$. If $V_1 \geq m$, then the m^{th} smallest

element of the original list lies in the left subtree, and we may locate it by applying Q_m to it. If $V_1 = m - 1$ then the key is the m^{th} smallest element and the process stops. Otherwise $V_1 < m - 1$, and the m^{th} smallest element is the $m - V_1 - 1$ smallest element in the right subtree, which we then locate by applying Q_{m-V_1-1} . Hence, letting $C_{n,m}$ denote the number of comparisons made by Q_m to process an n long list,

$$\begin{aligned} C_{n,m} &= 0 \quad \text{for } 0 \leq n \leq m - 1, \text{ and} \\ C_{n,m} &= n - 1 + C_{V_1,m} \mathbf{1}(V_1 \geq m) + C_{n-V_1-1,m-V_1-1} \mathbf{1}(V_1 < m - 1) \\ &= n - 1 + C_{V_1,m} + C_{n-V_1-1,m-V_1-1} \mathbf{1}(V_1 < m - 1) \quad \text{for } n \geq m, \end{aligned} \quad (21)$$

where we have dropped the indicator $\mathbf{1}(V_1 \geq m)$ on the second term, given that this case is handled by the boundary condition.

We begin the analysis of this more general case by bounding the expectation $E[C_{n,m}]$ of Q_m .

Lemma 2.6 *Let $C_{n,m}$ be the number of Quickselect comparisons for locating the m^{th} smallest element of a list of n distinct numbers. Then for all $m \geq 1$*

$$E[C_{n,m}] \leq 4n \quad \text{for all } n \geq 0. \quad (22)$$

Proof: Lemma 2.5 shows (22) holds for with m replaced by 1. Proceeding first by induction on m , set $a_{n,k} = E[C_{n,k}]$ and assume $a_{n,k} \leq 4n$ for all $1 \leq k \leq m - 1$ for some $m \geq 2$. By (21), and that V_1 is uniform over $\{0, \dots, n - 1\}$, we obtain

$$a_{n,m} = n - 1 + \frac{1}{n} \sum_{k=m}^{n-1} a_{k,m} + \frac{1}{n} \sum_{k=0}^{m-2} a_{n-k-1,m-k-1} \quad \text{for } n \geq m.$$

By the induction hypothesis,

$$a_{n,m} \leq n - 1 + \frac{1}{n} \left(\sum_{k=m}^{n-1} a_{k,m} + 4 \sum_{k=0}^{m-2} (n - k - 1) \right) = n - 1 + \frac{1}{n} \sum_{k=m}^{n-1} a_{k,m} + \frac{2}{n} (m - 1)(2n - m).$$

Now inducting on n to complete the inductive step in m , as $a_{n,m} = 0$ for $0 \leq n \leq m - 1$, we see (22) holds for $0 \leq n \leq m - 1$. Assuming (22) holds with n replaced by any k satisfying $0 \leq k \leq n - 1$ for some $n \geq m$, we obtain

$$\begin{aligned} a_{n,m} &\leq n - 1 + \frac{2}{n} (n(n - 1) - m(m - 1) + (m - 1)(2n - m)) \\ &= n - 1 + \frac{2}{n} (2m - 2m^2 - 3n + 2mn + n^2) \\ &\leq n - 1 + \frac{2}{n} (2m - 3n + \frac{3}{2}n^2) \leq n - 1 + 3n \leq 4n, \end{aligned}$$

where we have used

$$mn \leq \frac{1}{4}n^2 + m^2$$

to obtain the second inequality and that $m \leq n$ for the third. This inequality completes the inductive step in n and hence also in m , and the proof. \square

Letting $C_m(n; \mathbf{U}_1)$ be defined by rewriting (21), as (13) was derived from (11), we obtain

$$C_m(n; \mathbf{U}_1) = 0 \quad \text{for } 0 \leq n \leq m-1, \text{ and otherwise}$$

$$C_m(n; \mathbf{U}_1) = n-1 + C_m(\lfloor nU_1 \rfloor; \mathbf{U}_2) + C_{m-1-\lfloor nU_1 \rfloor}(n-1-\lfloor nU_1 \rfloor; \mathbf{U}_2)\mathbf{1}(\lfloor nU_1 \rfloor < m-1). \quad (23)$$

We next provide the following result needed for the proof of Theorem 1.2, which parallels Lemma 2.4 in the case $m = 1$.

Lemma 2.7 *For all $m \geq 2$ and $p \geq 1$*

$$f_p := E|C_m(p; \mathbf{U}_1) - C_m(p-1; \mathbf{U}_1)| \leq 2 + 16m.$$

Proof: As $C_m(p; \mathbf{U}_1) = 0$ for all $0 \leq p \leq m-1$ we may take $p \geq m$. By the basic recursion (21) we have

$$\begin{aligned} C_m(p; \mathbf{U}_1) - C_m(p-1; \mathbf{U}_1) &= 1 + C_m(\lfloor pU_1 \rfloor; \mathbf{U}_2) - C_m(\lfloor (p-1)U_1 \rfloor; \mathbf{U}_2) \\ &\quad + C_{m-1-\lfloor pU_1 \rfloor}(p-1-\lfloor pU_1 \rfloor; \mathbf{U}_2)\mathbf{1}(\lfloor pU_1 \rfloor < m-1) \\ &\quad - C_{m-1-\lfloor (p-1)U_1 \rfloor}(p-2-\lfloor (p-1)U_1 \rfloor; \mathbf{U}_2)\mathbf{1}(\lfloor (p-1)U_1 \rfloor < m-1) \\ &= 1 + (C_m(\lfloor pU_1 \rfloor; \mathbf{U}_2)\mathbf{1} - C_m(\lfloor (p-1)U_1 \rfloor; \mathbf{U}_2)) + R_1, \end{aligned}$$

where

$$\begin{aligned} R_1 &= C_{m-1-\lfloor pU_1 \rfloor}(p-1-\lfloor pU_1 \rfloor; \mathbf{U}_2)\mathbf{1}(\lfloor pU_1 \rfloor < m-1) \\ &\quad - C_{m-1-\lfloor (p-1)U_1 \rfloor}(p-2-\lfloor (p-1)U_1 \rfloor; \mathbf{U}_2)\mathbf{1}(\lfloor (p-1)U_1 \rfloor < m-1). \end{aligned}$$

Applying the triangle inequality and taking expectation,

$$f_p \leq 1 + E|C_m(\lfloor pU_1 \rfloor; \mathbf{U}_2)\mathbf{1} - C_m(\lfloor (p-1)U_1 \rfloor; \mathbf{U}_2)| + E|R_1|. \quad (24)$$

For the expectation in (24), by (20) we have

$$\begin{aligned} E|C_m(\lfloor pU_1 \rfloor; \mathbf{U}_2) - C_m(\lfloor (p-1)U_1 \rfloor; \mathbf{U}_2)| &= \sum_{k=1}^{p-1} E \left[|C_m(k, \mathbf{U}_2) - C_m(k, \mathbf{U}_2)| \mid U_1 \in \left[\frac{k-1}{p-1}, \frac{k}{p} \right) \right] P \left(U_1 \in \left[\frac{k-1}{p-1}, \frac{k}{p} \right) \right) \\ &\quad + \sum_{k=1}^{p-1} E \left[|C_m(k, \mathbf{U}_2) - C_m(k-1, \mathbf{U}_2)| \mid U_1 \in \left[\frac{k}{p}, \frac{k}{p-1} \right) \right] P \left(U_1 \in \left[\frac{k}{p}, \frac{k}{p-1} \right) \right) \\ &= \sum_{k=1}^{p-1} f_k P \left(U_1 \in \left[\frac{k}{p}, \frac{k}{p-1} \right) \right) = \frac{1}{p(p-1)} \sum_{k=1}^{p-1} k f_k. \end{aligned}$$

For the remainder term R_1 , applying Lemma 2.6 on the first term and using that $\lfloor pU_1 \rfloor \sim \mathcal{U}\{0, \dots, p-1\}$ yields

$$\begin{aligned} E[C_{m-1-\lfloor pU_1 \rfloor}(p-1-\lfloor pU_1 \rfloor; \mathbf{U}_2)\mathbf{1}(\lfloor pU_1 \rfloor < m-1)] &\leq \frac{4}{p} \sum_{k=0}^{m-2} (p-1-k) \\ &\leq \frac{4}{p} (p-1)(m-1) \leq 4m, \end{aligned}$$

and replacing p by $p - 1$ we see that the same bound holds for the expectation of the final R_1 term.

Substituting the bounds achieved into (24) we obtain

$$f_p \leq 1 + 8m + \frac{1}{p(p-1)} \sum_{k=1}^{p-1} k f_k \quad \text{for all } p \geq m. \quad (25)$$

As $f_p = 0$ for $1 \leq p \leq m - 1$ inequality (25) holds for all $p \geq 2$, and the conditions for invoking Lemma 2.3 with $c = 1 + 8m$ are satisfied, thus yielding the desired conclusion. \square

Proof of Theorem 1.2: Let $n \geq m$. From (23) and (3), letting $V_1 = \lfloor nU_1 \rfloor$,

$$\begin{aligned} W_n &= \frac{1}{n} C_m(n; \mathbf{U}_1) - 1 \\ &= \frac{1}{n} (n - 1 + C_m(V_1; \mathbf{U}_2) + C_{m-1-V_1}(n - 1 - V_1; \mathbf{U}_2) \mathbf{1}(V_1 < m - 1)) - 1 \\ &= \frac{1}{n} (C_m(V_1; \mathbf{U}_2) + C_{m-1-V_1}(n - 1 - V_1; \mathbf{U}_2) \mathbf{1}(V_1 < m - 1) - 1). \end{aligned} \quad (26)$$

We now construct a variable with the W_n^* distribution. As \mathbf{U}_1 and \mathbf{U}_2 are equidistributed, W'_n given by the first equality in (26) when substituting \mathbf{U}_2 in place of \mathbf{U}_1 has law $\mathcal{L}(W_n)$. Hence, by (1) with $\theta = 1$, letting

$$W_n^* = U_1(W'_n + 1) = \frac{1}{n} U_1 C_m(n; \mathbf{U}_2), \quad (27)$$

the pair (W_n, W_n^*) is a coupling of a variable with the W_n distribution to one with its Dickman \mathcal{D} -bias distribution. Applying consequence (8) of Theorem 1.3, we obtain

$$d_1(W_n, D) \leq \frac{2}{n} f_n \quad \text{where} \quad f_n = nE|W_n^* - W_n|. \quad (28)$$

Letting

$$e_n = E|U_1 C_m(n; \mathbf{U}_2) - C_m(\lfloor nU_1 \rfloor; \mathbf{U}_2)|,$$

in view of (26) and (27), and applying Lemma 2.6 to bound expectations for the form $E[C_{n,m}]$, we obtain

$$\begin{aligned} f_n &= nE|W_n^* - W_n| \\ &= E|U_1 C_m(n; \mathbf{U}_2) - C_m(V_1; \mathbf{U}_2) - C_{m-1-V_1}(n - 1 - V_1; \mathbf{U}_2) \mathbf{1}(V_1 < m - 1) + 1| \\ &\leq e_n + E|C_{m-1-V_1}(n - 1 - V_1; \mathbf{U}_2) \mathbf{1}(V_1 < m - 1)| + 1 \\ &\leq e_n + \frac{4}{n} \sum_{k=0}^{m-2} (n - 1 - k) + 1 \\ &\leq e_n + \frac{4}{n} (n - 1)(m - 1) + 1 = e_n + 4(m - 1) + 1 \leq e_n + 4m. \end{aligned} \quad (29)$$

To control e_n , invoke the basic recursion (23) to write

$$\begin{aligned} U_1 C_m(n; \mathbf{U}_2) &= U_1(n-1) + U_1 C_m(\lfloor nU_2 \rfloor; \mathbf{U}_3) \\ &\quad + U_1 C_{m-1-\lfloor nU_2 \rfloor}(n-1-\lfloor nU_2 \rfloor; \mathbf{U}_3) \mathbf{1}(\lfloor nU_2 \rfloor < m-1) \\ &= U_1(n-1) + U_1 C_m(\lfloor nU_2 \rfloor; \mathbf{U}_3) + R_1 \end{aligned}$$

where

$$R_1 = U_1 C_{m-1-\lfloor nU_2 \rfloor}(n-1-\lfloor nU_2 \rfloor; \mathbf{U}_3) \mathbf{1}(\lfloor nU_2 \rfloor < m-1),$$

and similarly,

$$\begin{aligned} C_m(\lfloor nU_1 \rfloor; \mathbf{U}_2) &= (\lfloor nU_1 \rfloor - 1) \mathbf{1}(\lfloor nU_1 \rfloor \geq m) + C_m(\lfloor \lfloor nU_1 \rfloor U_2 \rfloor; \mathbf{U}_3) + R_2 \\ &= (\lfloor nU_1 \rfloor - 1) + C_m(\lfloor \lfloor nU_1 \rfloor U_2 \rfloor; \mathbf{U}_3) + R_2 + R_3 \end{aligned}$$

where

$$R_2 = C_{m-1-\lfloor \lfloor nU_1 \rfloor U_2 \rfloor}(\lfloor nU_1 \rfloor - 1 - \lfloor \lfloor nU_1 \rfloor U_2 \rfloor; \mathbf{U}_3) \mathbf{1}(\lfloor \lfloor nU_1 \rfloor U_2 \rfloor < m-1, \lfloor nU_1 \rfloor \geq m),$$

and

$$R_3 = -(\lfloor nU_1 \rfloor - 1) \mathbf{1}(\lfloor nU_1 \rfloor \leq m-1).$$

Taking the expectation of the absolute difference, we obtain

$$\begin{aligned} e_n &= E|U_1 C_m(n; \mathbf{U}_2) - C_m(\lfloor nU_1 \rfloor; \mathbf{U}_2)| \\ &\leq E|U_1(n-1) - (\lfloor nU_1 \rfloor - 1)| + E|U_1 C_m(\lfloor nU_2 \rfloor; \mathbf{U}_3) \\ &\quad - C_m(\lfloor \lfloor nU_1 \rfloor U_2 \rfloor; \mathbf{U}_3)| + E|R_1| + E|R_2| + E|R_3| \\ &\leq 2 + E|U_1 C_m(\lfloor nU_2 \rfloor; \mathbf{U}_3) - C_m(\lfloor \lfloor nU_1 \rfloor U_2 \rfloor; \mathbf{U}_3)| + E|R_1| + E|R_2| + E|R_3|. \quad (30) \end{aligned}$$

Lemmas 2.2 and 2.7 yield

$$E|C_m(\lfloor \lfloor nU_1 \rfloor U_2 \rfloor; \mathbf{U}_3) - C_m(\lfloor \lfloor nU_2 \rfloor U_1 \rfloor; \mathbf{U}_3)| \leq 2 + 16m. \quad (31)$$

For the first remainder term R_1 , by Lemma 2.6, we have

$$\begin{aligned} E|R_1| &\leq \frac{1}{2} E[C_{m-1-\lfloor nU_2 \rfloor}(n-1-\lfloor nU_2 \rfloor; \mathbf{U}_3) \mathbf{1}(\lfloor nU_2 \rfloor \leq m-2)] \\ &\leq \frac{2}{n} \sum_{k=0}^{m-2} (n-1-k) = \frac{2}{n} (n-1)(m-1) \leq 2m. \quad (32) \end{aligned}$$

For R_2 , we condition on the event $\lfloor nU_1 \rfloor = k$ for $1 \leq k \leq n-1$, then further on $\lfloor kU_2 \rfloor = j$ for $0 \leq j \leq k-1$. We note the presence of $\lfloor nU_1 \rfloor \geq m$ in the indicator restricts $k \geq m \geq 2$ in this second step, where the values of j are all equally likely with probability $1/k$. Applying Lemma 2.6 then yields

$$\begin{aligned} E|R_2| &\leq E[C_{m-1-\lfloor \lfloor nU_1 \rfloor U_2 \rfloor}(\lfloor nU_1 \rfloor - 1 - \lfloor \lfloor nU_1 \rfloor U_2 \rfloor; \mathbf{U}_3) \mathbf{1}(\lfloor \lfloor nU_1 \rfloor U_2 \rfloor < m-1, \lfloor nU_1 \rfloor \geq m)] \\ &\leq \frac{4}{n} \sum_{k=m}^{n-1} \frac{1}{k} \sum_{j=0}^{m-2} (k-1-j) = \frac{4m}{n} \sum_{k=1}^{n-1} \frac{1}{k} (k-1) \leq \frac{4m}{n} (n-1) \leq 4m. \quad (33) \end{aligned}$$

As R_3 satisfies

$$E|R_3| = E|(\lfloor nU_1 \rfloor - 1)\mathbf{1}(\lfloor nU_1 \rfloor \leq m - 1)| \leq m, \quad (34)$$

substituting the bounds (31)-(34) into (30) yields

$$\begin{aligned} e_n &\leq 4 + 23m + E|U_1 C_m(\lfloor nU_2 \rfloor; \mathbf{U}_3) - C_m(\lfloor nU_2 \rfloor U_1; \mathbf{U}_3)| \\ &= 4 + 23m + \frac{1}{n} \sum_{k=0}^{n-1} e_k = 4 + 23m + \frac{1}{n} \sum_{k=m}^{n-1} e_k \end{aligned}$$

where the final inequality follows by noting that $C(k; \mathbf{U}_1) = 0$ for $k \leq m - 1$. Applying Lemma 2.1 yields

$$e_n \leq (4 + 23m) \log(ne/m),$$

and now from (29) we conclude

$$f_n \leq e_n + 4m = (4 + 23) \log(ne/m) + 4m,$$

and substitution into (28) yields the claim. \square

3 The Dickman transform and Stein equation

In this section we introduce and study the averaging operator (35) which is crucial to the solution of the Stein equation (45) for the generalized Dickman. After building the necessary tools, we present Theorem 3.1 and its proof, which is the key to demonstrating Theorem 1.3.

Let dv denote Lebesgue measure on $[0, \infty)$ and for $\theta > 0$ let $dv_\theta = \theta v^{\theta-1} dv$. For $a > 0$ and a v_θ integrable function $h : [0, a] \rightarrow \mathbb{R}$, define the averaging operator

$$A_x h = \frac{1}{x^\theta} \int_0^x h(u) \theta u^{\theta-1} du \quad \text{for } x \in (0, a]. \quad (35)$$

We adopt the convention that $A_0 h = h(0)$, and for notational simplicity suppress the dependence of A_x on θ . Note that

$$f(x) = A_x h \quad \text{for } x \geq 0 \text{ if and only if } (x/\theta)f'(x) + f(x) = h(x) \quad \text{a.e. on } [0, \infty). \quad (36)$$

The distribution of W^* as given in (1) can be written as a mixture of the laws of $U^{1/\theta}(w + 1)$ over realizations of $W + 1$, and thus has density

$$\theta x^{\theta-1} \int_x^\infty \frac{1}{v^\theta} dF(v), \quad x > 0$$

where F is the distribution of $W + 1$. In particular, W^* is absolutely continuous, and, recalling definition (5), the expectation $E[f'(W^*)]$ of an almost everywhere derivative of a function $f \in \text{Lip}_\alpha, \alpha \geq 0$ evaluated at W^* is well defined.

As the moment generating function of the generalized Dickman exists for all $t \in \mathbb{R}$ by Proposition 3 of [16], the expectation $E[h(D_\theta)]$ exists when h grows no faster than exponentially. In particular, the generalized Dickman distribution has moments of all orders. Using that $D_\theta =_d D_\theta^*$ in (1) yields

$$D_\theta =_d U^{1/\theta}(D+1) \quad \text{and hence} \quad E[D_\theta] = \frac{\theta}{\theta+1}(E[D_\theta] + 1), \quad \text{yielding} \quad E[D_\theta] = \theta. \quad (37)$$

Definition (1) leads to the following functional characterization of the relationship between W and W^* .

Lemma 3.1 *If W^* has the \mathcal{D}_θ bias distribution of W as in (1) then the equality and existence of both sides of*

$$E[h(W^*)] = E[A_{W+1}h] \quad (38)$$

hold for all functions h for which expectation of either side exists. If $f(x) = A_x h$ for any such h , then

$$E[(W^*/\theta)f'(W^*) + f(W^*)] = E[f(W+1)], \quad (39)$$

and all expressions in (38) and (39) are equal. If $E[W]$ exists, then (39) holds for all $f \in \bigcup_{\alpha \geq 0} \text{Lip}_\alpha$.

Proof: If the left hand side of (38) exists, then by definition (1) of W^* we have

$$\begin{aligned} E[h(W^*)] &= E[h(U^{1/\theta}(W+1))] = E \int_0^1 h(v(W+1)) \theta v^{\theta-1} dv \\ &= E \left[\frac{1}{(W+1)^\theta} \int_0^{W+1} h(v) \theta v^{\theta-1} dv \right] = E[A_{W+1}h]. \end{aligned}$$

If the right hand side of (38) exists, then we read the display above from right to left.

To show (39), note that when $f(x) = A_x h$ the right hand sides of (38) and (39) agree by definition of the averaging operator, and by the second equality in (36) their left hand sides also agree.

Lastly, as $E[W]$ exists, from (1), we see that $E[W^*]$ exists as well. If $f \in \text{Lip}_\alpha$ for $\alpha \geq 0$ then h as given in the second equality in (36) has at most linear growth. Hence $E[h(W^*)]$ also exists, and (39) holds for f . \square

For $\theta > 0$ and a differentiable function f , in view of (39), let

$$\mathbb{D}_\theta f(x) = (x/\theta)f'(x) + f(x) - f(x+1). \quad (40)$$

Note that when $f(x) = A_x g$, by (36) the equality (40) may be written

$$\mathbb{D}_\theta f(x) = g(x) - A_{x+1}g. \quad (41)$$

Lemma 3.2 *If W is a non-negative random variable, the following are equivalent:*

1.

$$W \sim \mathcal{D}_\theta.$$

2.

$$E[h(W)] = E[A_{W+1}h] \quad \text{for all } h \text{ such that } E[h(D_\theta)] \text{ exists.}$$

3.

$$E[\mathbb{D}_\theta f(W)] = 0 \quad \text{for all } f \in \bigcup_{\alpha \geq 0} \text{Lip}_\alpha.$$

Proof: When $W \sim \mathcal{D}_\theta$ then $W =_d W^*$ by Proposition 2 of [16], and now 2 follows by (38) of Lemma 3.1.

Now let 2 hold, and take $f \in \text{Lip}_1$ and $h(x)$ as given by the second equality in (36). Then $h(x)$ has at most linear growth, and as the D_θ has moments of all orders, $E[h(D_\theta)]$ exists. Using the definition of h for the first equality, the hypothesis for the second equality and that, by (36), $f(x) = A_x h$ for the third we have

$$E[(W/\theta)f'(W) + f(W)] = E[h(W)] = E[A_{W+1}h] = E[f(W+1)], \quad (42)$$

verifying $E[\mathbb{D}_\theta f(W)] = 0$ for all $f \in \text{Lip}_1$. Now given $f \in \text{Lip}_\alpha$ for $\alpha > 0$, the case $\alpha = 0$ being trivial, as $f/\alpha \in \text{Lip}_1$ we have $0 = E[\mathbb{D}_\theta(f(W)/\alpha)] = (1/\alpha)E[\mathbb{D}_\theta f(W)]$.

Now let 3 be satisfied and let $f \in \bigcup_{\alpha \geq 0} \text{Lip}_\alpha$. As $E[D_\theta]$ exists, the expectation in 3 of Lemma 3.2 does also. Let $h(x)$ again be given from f by (36). Then by the definition of h for the first equality, assumption 3 for the second and (38) of Lemma 3.1 for the third we have

$$E[h(W)] = E[(W/\theta)f'(W) + f(W)] = E[f(W+1)] = E[h(W^*)]. \quad (43)$$

For $0 < t \leq 1$ letting

$$f(x) = \begin{cases} 0 & x \leq t \\ 1 - \left(\frac{t}{x}\right)^\theta & x > t \end{cases}$$

we have $f \in \text{Lip}_{\theta/t}$ and $h(x) = \mathbf{1}(x > t)$, and from (43) we obtain

$$P(W \leq t) = 1 - P(W > t) = 1 - E[h(W)] = 1 - E[h(W^*)] = P(W^* \leq t).$$

By right continuity this equality holds also at zero. Likewise, for $t > 1$ we set

$$f(x) = \begin{cases} 1 & x \leq t \\ \left(\frac{t}{x}\right)^\theta & x > t \end{cases}$$

yielding $f \in \text{Lip}_{\theta/t}$ and $h(x) = \mathbf{1}(x \leq t)$, and therefore $P(W \leq t) = P(W^* \leq t)$ for all $t > 1$. Hence $W =_d W^*$, and W is generalized Dickman, showing 1. \square

With \mathbb{D}_θ given by (40), based on the characterization in 3 of Lemma 3.2 we obtain the differential Stein equation for \mathcal{D}_θ of the form

$$\mathbb{D}_\theta f(x) = h(x) - E[h(D_\theta)]. \quad (44)$$

By (41), we may also consider the equivalent integral equation

$$g(x) - A_{x+1}g = h(x) - E[h(D_\theta)]. \quad (45)$$

For $\alpha \geq 0$ and Lip_α defined in (5), suppressing θ for notational simplicity, let

$$\mathcal{H}_\alpha = \{h : [0, \infty) \rightarrow \mathbb{R} : h \in \text{Lip}_\alpha, E[h(D_\theta)] = 0\}. \quad (46)$$

The main goal of this section is to prove Theorem 3.1 which provides a smooth solution of (45); we show directly after its statement how Theorem 1.3 immediately follows from that result and a technical fact shown below in Lemma 3.4. For a real valued function g on $S \subset [0, \infty)$ define

$$\|g\|_S = \sup_{x \in S} |g(x)| \quad \text{and} \quad \|g\| = \sup_{x \in [0, \infty)} |g(x)|.$$

Theorem 3.1 *For every $\theta > 0$ and $h \in \mathcal{H}_1$ there exists a function $g \in \text{Lip}_{1+\theta}$ that solves (45).*

Proof of Theorem 1.3: Let $h \in \text{Lip}_1$ and note that $h - E[h(D_\theta)] \in \mathcal{H}_1$. Let g be the $\text{Lip}_{1+\theta}$ solution to (45) guaranteed by Theorem 3.1. As $E[W]$ exists and g has at most linear growth, $E[g(W)]$ exists. By 2 of Lemma 3.4 $A_{x+1}g \in \text{Lip}_1$ and hence $E[A_{W+1}g]$ exists as well.

Let (W, W^*) be any coupling of W to a variable W^* having its \mathcal{D}_θ -biased distribution. Now, from the Stein equation (45), followed by (38) of Lemma 3.1 and the mean value theorem, we obtain

$$\begin{aligned} |Eh(W) - Eh(D_\theta)| &= |E[g(W) - A_{W+1}g]| = |E[g(W) - g(W^*)]| \\ &\leq \|g'\| E|W^* - W| \leq (1 + \theta) E|W^* - W|, \end{aligned}$$

applying Theorem 3.1 for the final inequality. Taking the supremum over $h \in \text{Lip}_1$ on the left hand side and using (4) yields (8). Now the proof is completed by taking the infimum over all couplings of W and W^* on the right of (8), followed by (7). \square

We now derive some key properties that the averaging operator enjoys. For ease of notation we let

$$\rho = \theta/(\theta + 1).$$

Further, when considering a function, $A_x h$ for $x > 0$ for instance, as pure, or unevaluated, such as when writing membership in a function class such as in Lemma 3.3 below, we will write $A_\bullet h$.

Lemma 3.3 *Let $\theta > 0$ and $A_x h$ be given by (35). If $h \in \text{Lip}_\alpha$ for some $\alpha \geq 0$ then $A_x h$ exists for all $x \geq 0$ and $A_\bullet h \in \text{Lip}_{\alpha\rho}$.*

Proof: Let $h \in \text{Lip}_\alpha$. As h has at most linear growth, h is v_θ integrable on every interval $[0, x]$ for $x > 0$, hence $A_x h$ exists for all $x \geq 0$, recalling that $A_0 h = h(0)$. Let $h'(x)$ be any be a.e. derivative of $h(x)$. First assume that $h(0) = 0$. Using Fubini's theorem,

$$\begin{aligned} A_x h &= \frac{1}{x^\theta} \int_0^x h(u) \theta u^{\theta-1} du = \frac{1}{x^\theta} \int_0^x \int_0^u h'(v) \theta u^{\theta-1} dv du \\ &= \frac{1}{x^\theta} \int_0^x \int_v^x h'(v) \theta u^{\theta-1} du dv = \frac{1}{x^\theta} \int_0^x h'(v) [x^\theta - v^\theta] dv. \end{aligned}$$

Differentiating once we obtain

$$(A_x h)' = \frac{\theta}{x^{\theta+1}} \int_0^x h'(v) v^\theta dv \quad \text{for } x > 0, \quad (47)$$

and differentiation again yields

$$(A_x h)'' = \frac{\theta}{x^2} \left[x h'(x) - \frac{\theta+1}{x^\theta} \int_0^x v^\theta h'(v) dv \right] \quad \text{for } x > 0.$$

As $h \in \text{Lip}_\alpha$, we have $|h'(x)| \leq \alpha$ a.e. for all $x > 0$, and (47) then yields

$$|(A_x h)'| \leq \frac{\alpha\theta}{x^{\theta+1}} \int_0^x v^\theta dv = \frac{\alpha\theta}{\theta+1} = \alpha\rho \quad \text{for all } x > 0,$$

implying $A_x h \in \text{Lip}_{\alpha\rho}$. In addition, as $A_0 h = h(0) = 0$, we also now have

$$|A_x h| = |A_x h - A_0 h| \leq \alpha\rho|x - 0| \leq \alpha\rho x \quad \text{for all } x > 0,$$

which also holds trivially at $x = 0$.

In general, given any $h \in \text{Lip}_\alpha$ then $h_0(x) = h(x) - h(0)$ is in Lip_α and satisfies $h_0(0) = 0$, hence $A_x h_0 \in \text{Lip}_{\alpha\rho}$ and for any $\{x, y\} \subset [0, \infty)$,

$$|A_y h - A_x h| = |A_y h_0 - A_x h_0| \leq \alpha\rho|y - x|,$$

as $A_t h(0) = h(0)$ for all $t \geq 0$. □

For clarity, we state that the averaging operator A_{x+1} , which is A_x evaluated at $x+1$, is explicitly given by

$$A_{x+1} h = \frac{1}{(x+1)^\theta} \int_0^{x+1} h(u) \theta u^{\theta-1} du \quad \text{for all } x \geq 0. \quad (48)$$

Lemma 3.4 *With $\theta > 0$ let the averaging operator A_{x+1} be as in (48).*

1. *If h is a v_θ integrable function on $[0, a+1]$ then $\|A_{\bullet+1} h\|_{[0,a]} \leq \|h\|_{[0,a+1]}$.*
2. *If $h \in \text{Lip}_\alpha$ for some $\alpha \geq 0$ then $A_{x+1} h$ exists for all $x \geq 0$ and $A_{\bullet+1} h \in \text{Lip}_{\alpha\rho}$.*

Proof: For part 1,

$$\begin{aligned} \|A_{\bullet+1} h\|_{[0,a]} &= \sup_{x \in [0,a]} |A_{x+1}| \leq \sup_{x \in [0,a]} \frac{1}{(x+1)^\theta} \int_0^{x+1} |h(u)| \theta x^{\theta-1} du \\ &\leq \|h\|_{[0,a+1]} \frac{1}{(x+1)^\theta} \int_0^{x+1} \theta u^{\theta-1} du = \|h\|_{[0,a+1]}. \end{aligned}$$

For Part 2, Lemma 3.3 shows that if $h \in \text{Lip}_\alpha$ then $A_\bullet h$ exists is an element of $\text{Lip}_{\alpha\rho}$. Hence its shift $A_{\bullet+1} h$ also enjoys these properties. □

Lemma 3.5 *With \mathcal{H}_α as in (46), for any $\alpha \geq 0$*

$$\sup_{h \in \mathcal{H}_\alpha} |h(0)| = \alpha\theta. \quad (49)$$

Proof: It suffices to prove the claim when $h(0) \geq 0$, as $h \in \mathcal{H}_\alpha$ if and only if $-h \in \mathcal{H}_\alpha$. From (37) we see that $E[D_\theta] = \theta$, hence the function $h_\alpha(x) = \alpha(\theta - x)$ is an element of \mathcal{H}_α . As $h_\alpha(0) = \alpha\theta$, the supremum in (49) must be at least $\alpha\theta$. If $h \in \text{Lip}_\alpha$ with $h(0) = \alpha\theta + \epsilon$ for some $\epsilon > 0$, then

$$h(0) - h(x) \leq |h(x) - h(0)| \leq \alpha x \quad \text{hence} \quad h(x) \geq h(0) - \alpha x = \alpha\theta + \epsilon - \alpha x = h_\alpha(x) + \epsilon.$$

Hence $E[h(D_\theta)] \geq \epsilon$, implying $h \notin \mathcal{H}_\alpha$ and thus demonstrating that the supremum in (49) is at most $\alpha\theta$. \square

Lemma 3.6 *For a function $h \in \mathcal{H}_\alpha$ the iterates of the averaging operator A_{x+1} on h defined by*

$$A_{x+1}^0 h = h(x) \quad \text{and} \quad A_{x+1}^{n+1} h = A_{x+1}(A_{x+1}^n h) \quad \text{for } n \geq 0$$

exist for all $n \geq 0$, and for any $\alpha \geq 0$ and $n \geq 1$,

$$\mathcal{A}_{x+1}^n(\mathcal{H}_\alpha) \subset \mathcal{H}_{\alpha\rho^n}.$$

Proof: All claims of the lemma clearly hold for $n = 0$. If $h \in \mathcal{H}_\beta$ for some $\beta > 0$ then $h \in \text{Lip}_\beta$, hence $\mathcal{A}_{x+1}h$ exists for all $x \geq 0$ and $\mathcal{A}_{x+1}h \in \text{Lip}_{\beta\rho}$ by part 4 of Lemma 3.4. Additionally, as $E[h(D_\theta)] = 0$ we conclude that $E[\mathcal{A}_{D+1}h] = 0$ by part 2 of Lemma 3.2. Hence $\mathcal{A}_{x+1}h \in \mathcal{H}_{\beta\rho}$. Assuming the claims of the lemma hold for $n \geq 0$ and letting $\beta = \alpha\rho^n$ we see they also hold for $n + 1$. \square

For $h \in \mathcal{H}_\alpha$ for short let

$$h^{(\star n)}(x) := A_{x+1}^n h \quad \text{for all } n \geq 0 \text{ and all } x \geq 0.$$

Proof of Theorem 3.1: Let $\theta > 0$ and $h \in \mathcal{H}_1$. We first show that the sums

$$g(x) = \sum_{n \geq 0} h^{(\star n)}(x) \quad \text{and} \quad A_{x+1}g = \sum_{n \geq 1} h^{(\star n)}(x) \tag{50}$$

converge uniformly on compact intervals, and $g \in \text{Lip}_{1/(1-\rho)}$.

By Lemma 3.6 we have that $h^{(\star k)} \in \mathcal{H}_{\rho^k}$, implying that $h^{(\star k)} \in \text{Lip}_{\rho^k}$ and, by Lemma 3.5, that $|h^{(\star k)}(0)| \leq \theta\rho^k$. Hence, for any $a \geq 0$ we have $\|h^{(\star k)}\|_{[0,a]} \leq (a + \theta)\rho^k$. Write the partial sums of $g(x)$ in (50) as

$$g_n(x) = \sum_{k=0}^n h^{(\star k)}(x) \quad \text{for } n \geq 0.$$

As $h^{(\star k)} \in \text{Lip}_{\rho^k}$ we have $g_n \in \text{Lip}_{(1-\rho^{n+1})/(1-\rho)}$, and in particular g_n is continuous for all $n \geq 0$. For any $n \geq m \geq 0$ and any $a \geq 0$ we have

$$\|g_n - g_m\|_{[0,a]} \leq \sum_{k=m+1}^n \|h^{(\star k)}\|_{[0,a]} \leq (a + \theta) \sum_{k=m+1}^{\infty} \rho^k = (a + \theta) \frac{\rho^{m+1}}{1 - \rho}. \tag{51}$$

Hence $g_n, n \geq 0$ is a Cauchy sequence of continuous functions with respect to the supremum norm over $[0, a]$, and hence converges uniformly on $[0, a]$ to g , a continuous function, satisfying $g \in \text{Lip}_{1/(1-\rho)}$.

Applying 1 of Lemma 3.4 and letting n tend to infinity in (51) we obtain

$$\|A_{\bullet+1}g - \sum_{n=1}^m A_{\bullet+1}^n h\|_{[0,a]} = \|A_{\bullet+1}g - A_{\bullet+1}g_m\|_{[0,a]} \leq \|g - g_m\|_{[0,a+1]} \leq (a+1+\theta) \frac{\rho^{m+1}}{1-\rho}.$$

Now letting m tend to infinity shows the second equality in (50), with the sum converging uniformly over compact intervals.

It only remains to show g solves (45), which now follows from (50) via

$$g(x) - A_{x+1}g = \sum_{n \geq 0} A_{x+1}^n h - \sum_{n \geq 1} A_{x+1}^n h = A_{x+1}^0 h = h(x).$$

□

References

- [1] Arras, B., Mijoule, G., Poly, G. and Swan, Y. (2016) Distances between probability distributions via characteristic functions and biasing, <https://arxiv.org/abs/1605.06819>
- [2] Arratia, R., Goldstein, L. and Kochman, F. (2013) Size bias for one and all, <http://arxiv.org/abs/1308.2729>
- [3] Arratia, R., Barbour, A. and Tavaré, S. (2003) Logarithmic combinatorial structures: a probabilistic approach. EMS Monographs in Mathematics. European Mathematical Society (EMS), Zürich.
- [4] Barbour, A. and Nietlispach, B. (2011) Approximation by the Dickman distribution and quasi-logarithmic combinatorial structures, *Electron. J. Probab.* 16. no. 29, 880902.
- [5] Chen, L.H.Y., Goldstein, L. and Shao, Q.M. (2010) Normal approximation by Stein's method. Springer
- [6] Dadoun, B. and Neininger, R. (2014) A statistical view on exchanges in Quickselect. Proceedings of the Meeting on Analytic Algorithmics and Combinatorics, pp. 40-51.
- [7] Devroye, L. and Fawzi, O. (2010) "Simulating the Dickman distribution." *Statistics and Probability Letters*, 80.3 242-247.
- [8] Dickman, K. (1930). On the frequency of numbers containing prime factors of a certain relative magnitude, *Arkiv för Matematik, Astronomi och Fysik*, 22A (10): 114.
- [9] Goldstein, L. (2007) L^1 bounds in normal approximation, *Ann Prob*, 35, No. 5, 1888–1930
- [10] Goldstein, L. and Rinott, Y. (1996) Multivariate normal approximations by Stein's method and size bias couplings, *Journal of Applied Probability*, vol 33, pp. 1-17.
- [11] Hoare, C. A. R. (1961). Algorithm 65: Find, *Comm. ACM*, 4 (7): 321322.

- [12] Hsien-Kuei Hwang, and Tsung-Hsi Tsai (2002) Quickselect and the Dickman function. *Combinatorics, Probability & Computing*, 11.04 (2002): 353–371.
- [13] Rachev, S. T. (1991). Probability metrics and the stability of stochastic models, John Wiley & Son Ltd.
- [14] Pinsky, R. (2016) Natural Probabilistic Model on the Integers and its Relation to Dickman-Type Distributions and Buchstab’s Function. <https://arxiv.org/abs/1606.02965>
- [15] Pinsky, R. (2016) On the strange domain of attraction to generalized Dickman distributions for sums of independent random variables. <https://arxiv.org/abs/1611.07207>
- [16] Penrose, M. D., and Wade, A. R. (2004). Random minimal directed spanning trees and Dickman-type distributions, *Advances in Applied Probability*, 36(3), 691-714.
- [17] Stein, C. (1972) A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. Proc. Sixth Berkeley Symp. Math. Stat. Prob., (1972), pp. 583-602
- [18] Stein, C. (1986) Approximate Computation of Expectations, Institute of Mathematical Statistics, Hayward, CA.